Below 2 nm: Samsung 2 nm GAA, "Angstrom-Class" Branding, and Where ASML Sits in the New Lithography Regime

Douglas C. Youvan

doug@youvan.com

www.youvan.ai

January 2, 2026

The phrase "below 2 nm" sounds like a single finish line, but it is really a new operating regime: tighter patterning margins, heavier reliance on design-technology co-optimization, and a reshuffling of what "node leadership" even means. In this regime, Samsung's 2 nm GAA effort matters not only as a transistor story (nanosheets, power, variability), but as an industrial story about yield learning, volume credibility, and customer confidence. At the same time, the market's "1.x nm" labels—A16, 18A, 14A, 1.4-class—signal an industry that now uses node names as a generational shorthand rather than a direct geometric measurement. That shift changes how outsiders should interpret announcements, timelines, and comparative claims. Into this mix stands ASML, whose EUV tooling remains the bottleneck technology that quietly sets the tempo of advanced-node reality. The central question is not whether ASML matters, but how adoption bifurcates between extending low-NA EUV with multipatterning versus stepping into High-NA EUV as a new platform. This paper builds a disciplined map of the landscape: what "below 2 nm" concretely implies, why GAA is necessary but not sufficient, and how lithography platform choices propagate into cost, yield, and strategic power.

Keywords: Samsung 2nm, GAA, nanosheet, MBCFET, SF2, Exynos, yield learning, high volume manufacturing, node naming, angstrom era, ASML, EUV, low-NA EUV, High-NA EUV, multipatterning, DTCO, PPA, backside power delivery, foundry competition, semiconductor roadmap, lithography intensity.

A collaboration with GPT-5.2-Thinking. 45 pages. CC4.0.

Outline

1. Problem Statement: “Below 2 nm” as a New Regime
1.1 Why the phrase misleads: node labels versus physical dimensions
1.2 What actually tightens: variability, overlay, stochastic defects, and cost-per-good-die
1.3 The regime shift: scaling becomes a systems problem, not a transistor-only problem

2. Samsung’s 2 nm GAA Node: What Is Being Attempted
2.1 GAA/nanosheets in practice: what changes relative to FinFET
2.2 Why Samsung’s early GAA bet matters strategically
2.3 The “2 nm” deliverable: performance, power, density, and manufacturability targets
2.4 Yield learning as the real story: ramp mechanics and risk points

3. “Is Anyone Below 2 nm?” A Taxonomy of Claims
3.1 Three categories: shipping volume, risk production, and roadmapped branding
3.2 The angstrom-class naming turn: what 18A, A16, and 14A communicate
3.3 How to compare nodes without being fooled: proxy metrics and practical indicators
3.4 The limits of public information: why outsiders must reason probabilistically

4. Lithography as the Hidden Governor: Why ASML Becomes More Central, Not Less
4.1 EUV as a monopoly bottleneck: what “single supplier” means for the stack
4.2 Low-NA EUV extension: multipatterning, cost, and cycle-time tradeoffs
4.3 High-NA EUV: capability jump, engineering complexity, and adoption friction
4.4 “Lithography intensity”: why more steps can still increase ASML’s leverage
4.5 Installed base effects: service, upgrades, uptime, and the compounding advantage

5. Two Scaling Paths: Extend Low-NA or Jump to High-NA
5.1 Path A: Stretch low-NA with pattern splitting and computational correction
5.2 Path B: Use High-NA to reduce multipatterning, with new constraints elsewhere
5.3 How foundries choose: ROI, tool availability, learning curves, and customer pull
5.4 Consequences for design: layout rules, cell libraries, SRAM scaling, and PPA realism

6. GAA Is Necessary but Not Sufficient: The Full Stack of “Below 2 nm”
6.1 Interconnect and BEOL limits: resistance, capacitance, and routing congestion
6.2 Power delivery transitions: why backside power appears now
6.3 Process control: metrology, inspection, and stochastic defect management
6.4 Packaging and system scaling: advanced packaging as the “other half” of node leadership

7. Competitive Dynamics: Samsung, TSMC, Intel Through the ASML Lens
7.1 Differentiation strategies: platform timing, product anchoring, and customer signaling
7.2 The adoption fork: why one leader may delay High-NA while another pulls it in
7.3 Geopolitics and supply chain realities: export controls, capacity siting, and resilience
7.4 What “leadership” becomes: reliability, capacity, and ecosystem—beyond first claims

8. Implications for Customers: What to Ask Before Betting a Product on a Node
8.1 The customer due diligence checklist: yield, cycle time, defectivity, and roadmap credibility
8.2 Cost realism: masks, tool time, and the hidden tax of complexity
8.3 Risk posture: dual sourcing, design portability, and contingency planning
8.4 The “good enough” node: when a mature node can beat a bleeding-edge node

9. A Disciplined Way to Talk About “Below 2 nm” Without Hype
9.1 A vocabulary of claims: “branded,” “qualified,” “risk,” “HVM,” “profitable HVM”
9.2 Evidence hierarchy: what counts as signal versus PR
9.3 The public’s trap: why precision language is now a form of technical ethics

10. Conclusion: The Sub-2 nm Era as Lithography-Governed Systems Engineering
10.1 Restate the thesis: below 2 nm is a regime, not a number
10.2 Where Samsung fits: GAA ambition plus the real test of yield credibility
10.3 Where ASML sits: platform control, adoption timing, and compounding leverage
10.4 Forward work: case studies, comparative timelines, and a disciplined Reference section

References

Art

1. Problem Statement: “Below 2 nm” as a New Regime

The phrase “below 2 nm” is now doing the cultural work that “submicron” once did: it compresses a messy frontier into a single, memorable threshold. That is useful for headlines, investor decks, and casual comparison, but it is increasingly inaccurate as a technical description. What is actually happening is a transition in what sets the pace of progress. At older nodes, geometric shrink was a reasonably good summary of the main story: smaller features, more transistors per area, and a fairly direct connection between the “node number” and the physical scale of critical dimensions. At the leading edge today, that connection has weakened. “2 nm” is less a ruler reading than a generational label for a package of changes: new transistor geometry, new patterning constraints, new power-delivery strategies, heavier computational correction, and a different cost structure.

So the problem statement is not “who gets below 2 nm first?” The real question is: what are the necessary conditions for profitable, repeatable, high-yield manufacturing in a regime where physics noise, overlay limits, and defect statistics can dominate the economics? Put plainly: the “below 2 nm” era forces a reframing from scaling as a geometric event to scaling as a systems discipline, in which lithography, process control, design rules, interconnect, and packaging become co-equal constraints.

1.1 Why the phrase misleads: node labels versus physical dimensions

In modern foundry language, “2 nm” is not a literal claim about a single feature being 2 nanometers long. Multiple dimensions matter (gate pitch, metal pitch, contact pitch, fin/nanosheet dimensions, spacer thicknesses, line edge roughness), and no single one cleanly maps to the node label. The label functions as a generational marker: “this is the next major platform after the previous one.”

This matters because it changes how comparisons should be interpreted. If one company markets “1.8” or “1.6” or “1.4,” it does not automatically mean their transistors or wires are physically smaller in a directly comparable way. It may mean they have chosen a different naming scheme, are bundling different improvements into the “node,” or are highlighting a specific platform transition (for example, transistor architecture changes or power-delivery changes) that they feel merits a new label.

A disciplined reader should therefore treat node labels as *promises about an ecosystem*: design rules, IP availability, yield ramp confidence, expected performance-per-watt ranges, and cost targets. In other words, the label is closer to an “operating system version number” than a tape measure. If the user’s goal is to predict who will ship working chips, on what timeline, and at what cost, the correct questions are not “is it really 2 nm?” but “what are the pitches, what

patterning is required, what is the defectivity, and what is the yield maturity for meaningful die sizes?"

1.2 What actually tightens: variability, overlay, stochastic defects, and cost-per-good-die

When you push into the 2 nm-class and below, what tightens is not just geometry; it is the tolerance stack that determines whether the factory can repeatedly land patterns where the design expects them to land. Four coupled pressures become defining:

First, **variability** stops being a second-order annoyance and becomes first-order. At very small dimensions, tiny fluctuations in line width, edge roughness, spacer thickness, or nanosheet shape can cause large swings in device behavior. This is not only "random" in the everyday sense; it is variability rooted in materials granularity, process chemistry, and stochastic photon effects. The circuit designer experiences this as wider distributions of threshold voltage, drive current, leakage, and timing. The factory experiences it as an increasingly narrow window for "good-enough" process conditions.

Second, **overlay** becomes unforgiving. Overlay is the alignment accuracy between layers. If layer-to-layer placement errors approach the scale of the critical features being connected, you pay twice: yield loss from opens/shorts and performance loss from parasitic variation even on "functional" devices. Overlay is not a single knob; it reflects tool capability, wafer distortions, thermal history, and the cumulative effect of many process steps. Below 2 nm-class, the tolerance for overlay error becomes so tight that more of the manufacturing stack must be treated as a precision alignment problem, not merely a patterning problem.

Third, **stochastic defects** become economically loud. EUV lithography uses very short-wavelength light and relies on photon statistics, resist chemistry, and secondary electron effects. At smaller features, the number of photons contributing to a feature can be small enough that random fluctuations matter. That can generate missing or malformed features (classic "stochastic" failure modes). Even if these are rare per unit area, a large chip integrates an enormous area and an enormous number of features, so rare events become inevitable. This is where "defect density" is not an abstract metric; it directly converts into the probability that a die is sellable.

Fourth, **cost-per-good-die** becomes the only honest scoreboard. At advanced nodes, the cost is not merely the tool price; it is the total cycle-time and complexity tax: more masks, more steps, more inspection, more rework risk, and lower early yields. A node can be "technically real" yet commercially fragile if the cost-per-good-die is not competitive. This is also why "first silicon" is not the same as "viable manufacturing." The regime punishes prematurely optimistic narratives because every extra patterning step and every extra mask layer multiplies cost and schedule risk. Below 2 nm-class, the economic penalty for complexity rises faster than intuition suggests.

1.3 The regime shift: scaling becomes a systems problem, not a transistor-only problem

In the older shrink narrative, the transistor was the hero and everything else followed. In the below-2 nm regime, the transistor is necessary but not sufficient. You can "win" on transistor architecture and still lose on interconnect, power delivery, lithography throughput, defect control, or packaging constraints. The scaling frontier becomes a coupled system with multiple hard ceilings.

One way to see this is to list the interacting subsystems that now co-determine success:

- **Lithography platform and patterning strategy**: whether a foundry extends a given EUV numerical aperture with more multipatterning or adopts a new platform changes the entire cost and learning curve.
- **Process control and metrology/inspection**: as tolerances shrink, measurement and feedback must tighten. The factory becomes more like a closed-loop control system whose bandwidth and precision must increase.
- **Interconnect and BEOL scaling**: even if transistors get better, wires can get worse (higher resistance, worse capacitance, more congestion). Performance and energy can become wire-limited rather than device-limited.
- **Power delivery**: as power density rises and routing is constrained, the way power is delivered and distributed becomes a primary limiter. This is why architecture shifts in power delivery (including backside approaches in some roadmaps) appear as "node-defining" changes.
- **Design-technology co-optimization (DTCO)**: design rules, standard cell architecture, SRAM scaling, and placement/routing are now deeply entangled with process choices. The boundary between "design" and "process" is blurrier than it has ever been.

The conclusion is that "below 2 nm" is less a brag about miniature features and more a statement about systemic coordination. The winners will be those who can make the entire stack converge: lithography capability, yield learning, cost control, design enablement, and productization under schedule. In this regime, ASML's tool roadmap, Samsung's GAA yield maturity, and the industry's willingness to adopt new lithography platforms are not separate stories—they are the same story viewed from different layers of a single coupled system.

2. Samsung's 2 nm GAA Node: What Is Being Attempted

Samsung's 2 nm effort is best understood as an attempt to "make a new transistor family boring." That sounds like an insult, but it is exactly the point. At leading-edge nodes, the competitive prize is not merely inventing a better device; it is turning that device into a stable, manufacturable platform that yields predictably across many product types, design styles, and die sizes. Samsung's path is notable because it committed early to gate-all-around (GAA) nanosheet devices while much of the industry was still extracting life from FinFET. The bet is that the long-term scaling path is easier to sustain if you swallow the architectural discontinuity sooner, even if the near-term learning curve is steep.

At 2 nm-class, Samsung is attempting to demonstrate that GAA is not only viable, but productive: that it can deliver meaningful performance-per-watt gains and density gains while controlling variability and defectivity well enough to make the economics work. The node is also an industrial statement: Samsung must show it can ramp this platform at volumes that matter, with yields that convince customers their product schedules will not be held hostage by the factory's learning curve.

2.1 GAA/nanosheets in practice: what changes relative to FinFET

FinFETs controlled the channel by wrapping a gate around three sides of a fin. That was the key move that allowed scaling to continue when planar transistors lost electrostatic control. But FinFET geometry has a scaling ceiling: the fin can only be so narrow and so tall before mechanical, variability, and process limits become dominant. You also lose flexibility: effective channel width increases by adding fins, which forces quantization effects in device sizing and complicates circuit optimization.

GAA nanosheets change the geometry of control. Instead of a single fin with a gate on three sides, the channel is formed by one or more horizontal "sheets" (or "nanosheets") and the gate wraps around the channel on all sides. This provides stronger electrostatic control, which is valuable as gate lengths shrink and leakage pressures rise. It also provides a new and very practical knob: sheet width can be tuned to dial in drive current and leakage tradeoffs without needing to add whole new fins. In design terms, nanosheets promise more continuous device sizing and potentially better matching to library and SRAM constraints.

But the "what changes" is not only electrostatics. The manufacturing choreography changes in ways that are easy to underestimate:

- The stack requires forming multiple sheets with tight thickness and spacing control. Small deviations can shift threshold voltage, mobility, and leakage in ways that are harder to correct later.

- The gate stack must conformally wrap around the sheets with uniformity. Any voids, thickness variation, or interface defects show up directly as device variability.
- Contact formation and resistance become even more punishing. At these scales, a modest increase in contact resistance can erase the gains from better channel control.
- The device is more sensitive to line edge roughness and dimensional noise because the channels are thin; fractional changes are large percentage changes.

So, relative to FinFET, nanosheets offer more control and flexibility, but they also introduce new sources of variation and new integration steps that must be mastered. The "benefit" is real, but it is not free.

2.2 Why Samsung's early GAA bet matters strategically

Samsung's early GAA bet matters for one central reason: **the first mover advantage in a new device architecture is primarily a learning advantage.** When the industry moves from one dominant transistor architecture to the next, the winners are not simply those who show a slide claiming better PPA. The winners are those who accumulate manufacturing intuition: which process knobs actually matter, which failure modes dominate yields, how to make variability shrink over time, and how to build design rules that avoid catastrophic yield cliffs.

An "early" adoption has two potential strategic payoffs.

First, it can create a credibility narrative for customers who want a long runway. If a customer believes GAA is the inevitable future, they may prefer a foundry that has already fought the integration battles and stabilized its process. In that story, Samsung is trying to become the "experienced adult" in GAA while others are still "transitioning." Even if competitors catch up quickly, early learning can persist as a yield and ramp advantage in later generations.

Second, early adoption can shape the ecosystem. If Samsung can convince EDA, IP vendors, and large customers to co-design around its process assumptions, it can reduce friction for adoption. At advanced nodes, friction is expensive. The customer does not just buy transistors; they buy an entire enablement pipeline: libraries, SRAM compilers, analog IP, timing models, reliability models, sign-off rules, and packaging flows. A foundry that can stabilize these earlier can be more attractive even if raw density is not the highest.

Of course, there is a symmetrical risk: early adoption can also mean you absorb the pain first. If the learning curve is steep, initial yields can lag and customers may associate the foundry with schedule risk. That is why Samsung's 2 nm story cannot be judged by architecture choice alone; it must be judged by whether it can convert the bet into a stable manufacturing platform.

2.3 The “2 nm” deliverable: performance, power, density, and manufacturability targets

For customers, “2 nm” is a shorthand for a bundle of deliverables that must arrive together. A foundry can advertise “PPA gains,” but what customers actually require is a more granular set of guarantees:

- **Performance**: higher drive current at acceptable leakage, plus wiring and contact solutions that do not choke frequency scaling. It is common for device-level gains to be partially offset by interconnect and parasitic realities, so the deliverable must be system-level performance in real libraries, not just a single transistor plot.
- **Power**: improved energy efficiency, especially at iso-performance. Advanced nodes are justified as much by power savings as by speed. But power improvements must hold across PVT corners and across realistic workloads, not just at a marketing point.
- **Density**: real scaling in standard-cell area and SRAM bitcell area, not only in a “theoretical transistor count.” Density is where node labels often mislead, because density depends on design rules and routing constraints, not just transistor geometry.
- **Manufacturability**: stable design rules, predictable yield by die size, acceptable defect density, acceptable cycle time, and a path to cost-per-good-die that is competitive with the “next best alternative” node (often a mature node with aggressive libraries and advanced packaging).

The hard truth is that customers do not buy “2 nm.” They buy **schedule confidence**. That is why manufacturability is not a footnote; it is the center. Samsung’s deliverable is therefore not merely a 2 nm GAA transistor. It is a foundry platform that can support product launches without heroic, one-off engineering interventions.

There is also a subtle but crucial sub-deliverable: **the node must be designable.** The rules must not be so restrictive that only one internal chip team can make it work. A healthy foundry node is one that third parties can use without suffering an invisible tax of respins, margins, and late-stage surprises.

2.4 Yield learning as the real story: ramp mechanics and risk points

At the frontier, yield is the only metric that forces truth. You can claim almost anything in a slide deck; you cannot hide the cost of bad yield when wafers are expensive and customers demand volume.

Yield learning at 2 nm-class has a recognizable arc, and Samsung’s ramp will likely be judged against it:

- **Early silicon**: functional chips appear, but yields are low and variable. Many issues are "first-of-kind" integration problems: patterning defects, contact resistance excursions, random catastrophic defects, and systematic layout sensitivity.
- **Stabilization**: the factory identifies dominant defect modes, improves process windows, and tightens control. Yield starts to climb, but often unevenly across different design styles and die sizes.
- **Generalization**: the critical phase. The process must become robust across a wide variety of customer designs, not just one anchor SoC. The foundry must prove it can produce many chips, not just one chip.
- **Cost convergence**: yield improves enough that cost-per-good-die becomes plausible for commercial products. At this phase, cycle time and throughput become just as important as defectivity.

The risk points are likewise predictable, and they tend to hide in the seams between subsystems:

- **Overlay and patterning interactions**: when alignment errors and stochastic pattern defects combine, you can see sharp yield cliffs that depend on layout and routing patterns. The same process can look "fine" on one design and disastrous on another.
- **Contact and via resistance**: at tiny dimensions, resistance is not a nuisance; it is a product killer. A process that meets device targets but misses contact targets can underperform or fail power/timing closure at scale.
- **Variability-driven parametric yield loss**: even if chips "work," distributions can be too wide, causing bins to shift downward and revenue to drop. Parametric yield is often the silent killer: the chip passes tests but not at the desired frequency or voltage.
- **Inspection and metrology capacity**: advanced nodes require more measurement, more inspection, and tighter feedback loops. If measurement cannot keep up, the factory loses visibility, and yield learning slows.
- **Design-rule maturity**: rules that are too aggressive can generate late-stage yield surprises; rules that are too conservative can destroy density and performance, undermining the node's value proposition.

So the real Samsung question is not "can they make a 2 nm GAA transistor?" It is: can they drive yield learning through the generalization phase fast enough that external customers trust the node for flagship products? In the below-2 nm regime, trust is manufactured: it is the emergent

output of stable yields, repeatable cycle time, and credible schedules. The node's fate will be decided less by the brilliance of the transistor and more by the discipline of the ramp.

3. "Is Anyone Below 2 nm?" A Taxonomy of Claims

The question "is anyone below 2 nm?" sounds crisp, but it is usually asking several different things at once. It might be asking about physics ("are features literally smaller?"), manufacturing ("is this in high-volume?"), or market positioning ("has someone announced a 1.x node?"). In the current era, those questions do not align cleanly. Node names have become generational brand markers, while the real differentiators live in a web of pitches, patterning strategies, yield maturity, and design enablement.

This section proposes a disciplined taxonomy so the phrase "below 2 nm" can be used without confusion. The basic move is to separate claims by what they actually assert: production reality, production intent, or roadmapped identity. Once you do that, you can interpret "1.8," "1.6," "1.4," and "angstrom-class" not as ruler readings, but as signals about platform transitions, investment timing, and competitive narrative.

3.1 Three categories: shipping volume, risk production, and roadmapped branding

The first step is to classify any "below 2 nm" statement into one of three categories. These categories are not moral judgments. They are simply different kinds of truth claims.

Shipping volume means the node is producing sellable chips at a scale that matters commercially. This implies (a) stable yields that support large die sizes and real product margins, (b) predictable cycle time, and (c) a supply chain and packaging path capable of meeting customer demand. The key word is "repeatable." Shipping volume is not "we made a chip." It is "we can supply the market."

Risk production means the node is real in silicon and is entering an early manufacturing phase where customers can take meaningful design bets, but with recognized risk. Risk production can include pilot lines, limited early runs, engineering samples, and early customer engagements. Yields may still be in flux, design rules may still be evolving, and the foundry is still learning which failure modes dominate across diverse designs. Risk production is a commitment, but it is not yet proof of economic maturity.

Roadmapped branding is the public declaration that a company intends to have a node generation identified as 1.x nm or "angstrom-class." This is the most common source of confusion, because the market naturally hears "1.4 nm" as "smaller than 2 nm," while the company may be signaling a platform step: a future integration of new lithography, a new power-delivery scheme, a revised transistor stack, or a denser set of design rules. Roadmaps are

valuable. They shape ecosystem planning. But they are not evidence of current manufacturability.

A single company can simultaneously occupy multiple categories: it can ship one node in volume, run the next in risk production, and roadmap the one after that. The user's question "below 2 nm" becomes meaningful only when we pin down which category is being asked.

3.2 The angstrom-class naming turn: what 18A, A16, and 14A communicate

The emergence of "angstrom-class" labels is not an accident. It is a response to two realities.

First, the nanometer label has lost literal meaning. "2 nm" is not a single dimension, and "1.4 nm" is not a single dimension. The label is now primarily a generational marker. Moving to angstrom-style labels is a way to signal a new era while subtly acknowledging that the labels are conventions, not measurements.

Second, the market needs a vocabulary for platform transitions that are no longer just "shrink." Names like 18A, A16, and 14A communicate several things at once:

- A claim of being in the next generational cohort, beyond "2 nm-class."
- A claim that the company has a path to continue scaling where older playbooks are strained.
- A positioning move: "we are not stuck optimizing yesterday; we have a forward platform."

But the most important part is what they do not communicate. They do not, by themselves, tell you the real pitches, the patterning method, the yield maturity, or the cost structure. They are flags planted on a roadmap.

A useful way to interpret these names is: **they are coordination labels**. They help align internal engineering teams, tool suppliers, EDA/IP ecosystems, and customers around a timeline. They are also a negotiation instrument: they influence who commits to what tool purchases, what IP development, and what early product anchors. In that sense, angstrom-class branding is partly technical and partly industrial theater. Not "fake," but "not self-verifying."

3.3 How to compare nodes without being fooled: proxy metrics and practical indicators

If node labels are not literal, how do you compare nodes responsibly? You use proxy metrics that are harder to game and more directly tied to manufacturability and product outcomes. The goal is not perfect precision; the goal is to reduce susceptibility to branding.

A practical comparison toolkit includes:

Pitch proxies
Instead of trusting the node name, look for the reported or implied gate pitch and metal pitch, or any credible density metric tied to standard cells and SRAM. The exact numbers may not always be public, but the discipline is to seek "pitches and rules," not "node names."

Design enablement indicators
Ask whether standard-cell libraries, SRAM compilers, analog IP, and interface IP are mature and broadly available. A node that exists only as a heroic internal demo is different from a node that has a thriving external ecosystem. Enablement maturity is a real, often underappreciated leading indicator of commercialization.

Yield maturity signals
Yield is often not disclosed directly, but you can infer maturity from: the kinds of products announced, the die sizes involved, and whether the node is being used for flagship-volume parts or niche demonstrators. A foundry confident in yield tends to court bigger, riskier designs and larger volumes.

Cycle-time and complexity signals
More masks, more patterning steps, and heavier inspection requirements increase cycle time and cost. When you see a node whose story relies heavily on multipatterning extensions, interpret that as a plausible choice, but one that carries an economic and schedule tax. The relevant question becomes: is the platform economically competitive at the product level?

Performance-per-watt in real design contexts
"Transistor performance" is not "chip performance." Look for claims that appear in the context of real libraries, realistic voltages, and practical frequency targets. The more a claim is anchored to an actual product and its constraints, the less it resembles marketing vapor.

Customer behavior
This is the quietest but often strongest indicator. Who is committing flagship designs? Who is dual-sourcing? Who is hedging? Customer behavior aggregates a lot of private information: yield confidence, schedule risk, and commercial terms.

Using these proxies does not guarantee truth. But it shifts the discussion from slogans to signals that correlate with real manufacturing outcomes.

3.4 The limits of public information: why outsiders must reason probabilistically

Even with a disciplined proxy set, outsiders face a structural limitation: many of the most important parameters are proprietary. Foundries do not publish full pitch tables, defect Pareto charts, parametric yield distributions, or the hard details of what was required to make a

specific product close timing and yield. Customers under NDA also avoid discussing painful details. The public record is therefore incomplete by design.

That means a responsible analyst must reason probabilistically. Instead of concluding "Company X is ahead," the correct posture is usually: "Given the available signals, Company X appears to be ahead on platform A, while Company Y appears to be ahead on ramp maturity, and Company Z appears to be ahead on ecosystem readiness."

This probabilistic posture is not a weakness. It is epistemic hygiene. The alternative is being manipulated by whichever narrative is loudest.

A useful rule in this regime is to separate two layers of belief:

- **Belief about capability**: "They can likely make functional silicon at this generation."
- **Belief about commercial reality**: "They can likely supply profitable volume with predictable schedules."

Capability typically precedes commercial reality by a meaningful interval, and the interval can vary widely. The below-2 nm era amplifies that gap because the tolerance stack is tighter and the economic penalties for immaturity are harsher.

So, when someone asks "is anyone below 2 nm?", the disciplined answer is: yes, in brand labels and roadmaps; sometimes, in risk production; and only rarely, in the sense that matters most—profitable, repeatable shipping volume. The taxonomy turns a confusing question into a precise one, and precision is the first defense against being fooled.

4. Lithography as the Hidden Governor: Why ASML Becomes More Central, Not Less

If you want a single "governor" variable for advanced-node reality, it is not transistor architecture by itself. It is lithography capacity, capability, and learning rate. The reason is simple: the transistor, the interconnect, and the design rules are all downstream of what can be patterned reliably, repeatedly, and economically. When you cross into 2 nm-class and below, that dependence tightens. The patterning problem becomes more fragile; the tolerance stack becomes narrower; and the number of places where "one tiny defect becomes a dead die" increases. In that regime, the company that supplies the indispensable patterning tool does not become peripheral. It becomes more central.

ASML's centrality is sometimes misunderstood as a mere "EUV monopoly" talking point. But the deeper point is that EUV is not a commodity machine. It is a platform that determines what kinds of process flows are even plausible. It dictates how many masks are required, what kinds of computational corrections become mandatory, how inspection is prioritized, and how quickly

yield learning can proceed. Lithography doesn't just print patterns; it sculpts the entire factory's economics and the entire industry's cadence.

4.1 EUV as a monopoly bottleneck: what "single supplier" means for the stack

"Single supplier" is not only about pricing power. It is about systemic dependency. EUV is an interlocking ecosystem of source power, optics, masks, resists, metrology, and computational correction. In practice, the EUV scanner is the anchor around which the rest of the patterning ecosystem stabilizes. When one company supplies that anchor, several consequences follow.

First, **synchronization risk** becomes real. The entire leading-edge industry's timing is partially synchronized to ASML's product roadmap: when a new scanner generation reaches maturity, what throughput it can sustain, what uptime it can achieve, and how quickly it can be deployed in volume. If the platform's progress slows, the industry's progress slows. If the platform's progress accelerates, the industry gains headroom.

Second, **capital allocation becomes path-dependent**. Foundries and IDMs must commit billions to fabs years in advance. Those commitments depend on credible assumptions about lithography availability and productivity. In a single-supplier world, the supply of the critical machine becomes a gating function for how many wafers-per-month can be realistically planned.

Third, **ecosystem alignment becomes concentrated**. Mask shops, resist suppliers, inspection vendors, and EDA teams all tune their work to the realities of the scanner: defect modes, overlay performance, dose constraints, and throughput. In a multi-supplier world, competition can create divergent ecosystems. In a single-supplier world, the ecosystem converges around one platform's physics.

Finally, **the bottleneck becomes strategic**. Export controls, allocation decisions, long lead times, and service support become part of geopolitics and industrial policy. "Single supplier" means the world's most advanced patterning capability is not only an engineering constraint; it is an international lever.

The key point: monopoly here is not just market structure. It is architectural centrality in the technology stack.

4.2 Low-NA EUV extension: multipatterning, cost, and cycle-time tradeoffs

Low-NA EUV (the established EUV generation) still has runway. Extending it is not irrational; it is often economically motivated. The extension strategy is straightforward in concept: if one exposure cannot print the desired pattern with sufficient fidelity and yield, you split the pattern into multiple exposures and recombine them through etch and deposition steps. That is multipatterning, now performed in an EUV context as well as with older DUV methods.

This strategy has a clear advantage: it avoids the immediate leap to a new lithography platform. It uses a known tool ecosystem, known fab integration experience, and known supply chains. If the foundry can manage the complexity, it can delay the risk and cost of adopting a new scanner generation.

But the tradeoffs are severe, and below 2 nm they become defining:

- **Mask count rises.** Each additional pattern split often requires additional masks. Masks are expensive, and mask turnaround time can constrain development schedules.
- **Process steps multiply.** More exposures typically mean more etch/deposition cycles and more opportunities for defect introduction.
- **Overlay pressure intensifies.** Multiple patterning steps require precise layer-to-layer alignment. Overlay errors compound.
- **Cycle time increases.** More steps means longer wafer cycle times, which means higher work-in-progress, more inventory cost, and slower learning loops.
- **Yield sensitivity increases.** When you add steps, you add opportunities for both random and systematic defects. Even if each step is "pretty good," the compounded probability can be punishing.

So low-NA extension is a conscious decision to accept a complexity tax in exchange for avoiding a platform jump. In some product contexts, that trade can be rational. But it moves the battle from "can we pattern this?" to "can we pattern this profitably and on schedule?"

4.3 High-NA EUV: capability jump, engineering complexity, and adoption friction

High-NA EUV represents a different kind of scaling move: instead of splitting patterns into more steps, you improve the fundamental imaging capability so that more of the desired features can be patterned in fewer exposures. In principle, this can reduce multipatterning, reduce mask count, reduce cycle time, and improve fidelity for the most challenging layers.

However, High-NA is not simply "a better EUV machine." It is a platform jump with real adoption friction:

- **Tool complexity and footprint.** High-NA systems are larger, more complex, and more demanding in installation and fab integration.
- **New process learning.** A new imaging regime changes resist behavior, stochastic defect behavior, and the optimization of dose and focus.
- **Mask and pellicle constraints.** Advanced masks and their defect management become even more critical.

- **Ecosystem readiness.** The entire supporting ecosystem—resists, metrology, inspection, computational lithography—must mature in parallel.
- **Cost and ROI uncertainty.** High-NA tools are extraordinarily expensive. The foundry must believe the ROI is positive relative to simply doing more multipatterning on existing tools.

This is why adoption can be uneven. One company may adopt High-NA early because it wants a platform narrative, needs the capability for its chosen design rules, or has a strategic preference to reduce multipatterning complexity. Another may delay High-NA because it believes low-NA extension can reach the next node profitably, or because it wants the ecosystem to mature before taking platform risk.

The important point is that both strategies can be technically coherent. The difference is economic posture and risk tolerance. High-NA is a capability jump, but it demands a willingness to absorb early learning costs.

4.4 "Lithography intensity": why more steps can still increase ASML's leverage

A subtle but powerful concept in this era is **lithography intensity**: how much lithography tool time is required per wafer for a given node and product mix. Even if High-NA adoption is delayed, lithography intensity can rise because the patterning becomes harder. When patterns must be split, exposures increase. When yield learning demands more process monitoring, rework and measurement cycles can rise. When schedules tighten, more capacity is demanded to maintain throughput.

This matters because it means ASML can gain leverage even without a platform jump. If a node requires more exposures per wafer, then the same wafer output requires more scanner capacity. If the industry's "work per wafer" rises, the demand for tools and the value of uptime rise.

Lithography intensity also amplifies the value of incremental improvements: a few percentage points of throughput or uptime can translate into large effective capacity gains at the fab level. In a high-intensity regime, marginal improvements in scanner performance have outsized economic impact. That is exactly the kind of environment where the platform supplier's influence compounds.

So the common intuition—"if customers delay High-NA, ASML loses"—is incomplete. Delaying High-NA may shift where revenue appears (more low-NA tools, more services, more upgrades), but the structural dependence on ASML remains, and can even strengthen as complexity rises.

4.5 Installed base effects: service, upgrades, uptime, and the compounding advantage

The EUV story is not only about selling machines; it is about operating a fleet. As EUV becomes more central, the installed base becomes a compounding advantage for ASML and a compounding dependency for the industry.

Several effects stack together:

- **Service as a strategic layer.** Keeping scanners running at high uptime is not a routine maintenance issue; it is a leading-edge competitiveness issue. Differences in uptime translate directly into wafers shipped and learning speed.
- **Upgrades extend platform life.** ASML can improve productivity and capability through incremental upgrades and new subsystems. This allows fabs to get more value from existing capital, and it allows ASML to monetize the installed base beyond initial sales.
- **Operational learning compounds.** A large installed base generates data about failure modes, optimization strategies, and best practices. That data, combined with deep service relationships, can translate into faster improvements and better performance in the field.
- **Customer lock-in becomes structural.** Once a foundry's entire process flow and yield learning is built around a given scanner platform, switching costs are not merely financial; they are epistemic. You would have to relearn your factory. Even if another supplier existed, the inertia would be enormous.

In the below-2 nm regime, the "factory as a learning system" viewpoint becomes essential. The faster the learning loop, the faster yield improves, and the sooner cost-per-good-die becomes competitive. Lithography tools sit at the center of that loop because they shape defect modes, overlay constraints, and throughput. That is why ASML's installed base and service capability are not peripheral details. They are central to how quickly the entire industry can progress.

The bottom line is that lithography is the hidden governor because it sets the feasible space: what patterns can be printed, how many steps are required, how fast wafers move, and how rapidly the factory learns. In a world where the tolerance stack is tightening and node labels are increasingly conventional, the company that controls the essential lithography platform becomes more central, not less—because the entire definition of "advanced node reality" is increasingly "lithography-constrained systems engineering."

5. Two Scaling Paths: Extend Low-NA or Jump to High-NA

In the below-2 nm regime, “scaling” is no longer a single track. It is a choice between two coherent, competing strategies for managing the same physical reality: printing finer, denser patterns with acceptable yield and cost. One strategy extends the existing EUV platform (low-NA) by accepting more process complexity: you split patterns into multiple exposures, use more computational correction, and rely on tight overlay and defect control to keep yield afloat. The other strategy adopts a new platform (High-NA EUV) to regain patterning headroom, aiming to reduce multipatterning and simplify the most painful layers, at the price of a major tool and ecosystem transition.

Neither strategy is “obviously superior” in the abstract. Each is a trade between capital cost, learning risk, cycle time, design constraints, and the ability to sustain a predictable manufacturing rhythm. The choice is also time-dependent: what is rational at the start of a node can become irrational later as volumes rise, product mixes diversify, and the ecosystem matures. This section treats the two paths as legitimate engineering-economic strategies, then asks what determines which path a foundry will actually take.

5.1 Path A: Stretch low-NA with pattern splitting and computational correction

Path A is the extension strategy: keep using the established EUV platform and push it further by adding sophistication around it. The core idea is to accept that a single exposure will not always print the required features with enough fidelity and process window, and therefore to split the target pattern into components that are individually printable. These components are then recombined through multiple lithography-etch cycles.

In practice, this extension strategy is not just “more masks.” It is a stack of compensating technologies:

- **Pattern splitting and decomposition.** Layouts are decomposed into multiple masks such that each mask prints a subset of features with improved printability. This requires careful optimization to avoid creating hard-to-align interactions between the splits.
- **Computational lithography (OPC and beyond).** Optical proximity correction and model-based corrections become more intense because the margin for error shrinks. The design shape that is written on the mask can be far from the intended final shape on silicon.
- **Process window management.** The foundry must identify dose/focus windows that are simultaneously manufacturable and stable. As features shrink, the process window can collapse; Path A is about keeping it open through decomposition and correction.

- **Overlay discipline.** Because features are assembled from multiple exposures, overlay is now a yield-critical variable. The foundry's ability to keep overlay tight across a full wafer and across many steps becomes a defining capability.
- **Defect and stochastic management.** More steps mean more opportunities for defects. Path A requires aggressive inspection and rapid feedback loops so that defect sources are identified and suppressed before they dominate yield.

The appeal of Path A is immediate: it uses a familiar tool ecosystem and leverages existing fab investments. It also spreads risk: instead of betting on a single new platform with unknown learning curves, you accept incremental complexity in a platform you already know how to run at high uptime.

But the costs are structural:

- **Cycle time grows.** More steps means wafers take longer to traverse the fab. That slows learning, increases WIP, and can degrade responsiveness.
- **Mask costs rise and schedules tighten.** Mask sets become more expensive, and mask turnaround can become a schedule bottleneck.
- **Yield becomes more layout-sensitive.** The same process can behave differently depending on routing patterns, density contexts, and how the pattern splits interact. That increases the burden on design rules and sign-off.
- **The "complexity tax" compounds.** Each additional step multiplies opportunities for small excursions. Even if every step is "good," the compounded yield can still be painful.

Path A is therefore a pragmatic extension strategy, but one that gradually shifts the scaling problem from "can we print this?" to "can we afford to print this at scale?"

5.2 Path B: Use High-NA to reduce multipatterning, with new constraints elsewhere

Path B is the platform-jump strategy: adopt High-NA EUV to regain imaging capability so the most challenging layers require fewer splits and fewer exposures. Conceptually, the advantage is clear: fewer pattern splits can mean fewer masks, fewer steps, less overlay accumulation, and potentially better fidelity for extremely tight pitches.

However, Path B does not eliminate constraints. It moves them and introduces new ones.

First, **tool integration and availability** become immediate issues. A new platform is not only expensive; it is scarce. You must secure tools, install them, qualify them, and run them with high uptime. During early adoption, throughput and uptime can lag mature low-NA systems, which can partially erase the theoretical simplification benefits.

Second, **ecosystem maturity becomes a gating function.** High-NA changes imaging conditions, which can affect resist performance, stochastic defect behavior, and mask requirements. The foundry is now dependent on simultaneous progress in resists, masks, inspection, and computational models. If any of these lag, the platform's theoretical advantages can remain unrealized.

Third, **new design constraints can appear.** A platform jump can lead to new design rules, new forbidden patterns, and new sensitivities. Even if multipatterning is reduced, the process may impose its own restrictions that affect standard cells and SRAM.

Fourth, **the "elsewhere" constraints become visible.** If lithography becomes less of a bottleneck on certain layers, other bottlenecks can become dominant: contact resistance, interconnect RC, electromigration margins, thermal issues, and power delivery. Path B may make patterning easier while exposing that the rest of the system is now the limiter.

So Path B is not "free scaling." It is a rebalancing: you spend capital and accept platform risk to reduce complexity in patterning, betting that the overall system becomes simpler, faster, and more cost-effective at scale.

5.3 How foundries choose: ROI, tool availability, learning curves, and customer pull

The choice between Path A and Path B is not made purely by engineers and not purely by finance. It is made at the intersection of economics, tool supply, learning dynamics, and customer commitments.

A few decision variables tend to dominate:

ROI and cost-per-good-die
The practical goal is not "minimum nm." It is competitive cost-per-good-die at a performance-per-watt level customers will pay for. If low-NA extension can hit that target, a foundry can rationally delay High-NA. If the complexity tax becomes too large, High-NA can become the cheaper path despite its tool cost.

Tool availability and ramp timing
Even if High-NA is technically attractive, it may not be available in the quantities needed for a broad ramp. A foundry that cannot secure enough tools cannot bet its volume node on them. Conversely, a foundry that does secure early tools can use them to differentiate, even if early economics are harsh.

Learning curve risk
Foundries fear schedule surprises. A platform jump can introduce unknown yield limiters that take time to discover. If a foundry believes it can manage those unknowns, it may jump. If it believes those unknowns threaten customer schedules, it may extend low-NA.

Customer pull and anchor products
The single most important factor can be the existence of "anchor" products: large, high-profile chips whose success justifies the risk. When a customer commits a flagship product to a node, the foundry can align resources and accelerate learning. Customer pull can also be conditional: customers may demand a path to certain performance-per-watt targets that effectively forces a platform choice.

Ecosystem coordination
EDA flows, IP readiness, packaging options, and test infrastructure all matter. A foundry may choose the path that best fits what its ecosystem can support on the required timeline.

A useful way to summarize the decision: foundries choose the path that minimizes the probability-weighted sum of (a) schedule failure, (b) yield failure, and (c) cost failure, while maximizing customer confidence. This is why two rational foundries can choose different paths at the same time.

5.4 Consequences for design: layout rules, cell libraries, SRAM scaling, and PPA realism

The choice of lithography path is not confined to the fab. It propagates into design, sometimes brutally.

Layout rules and forbidden patterns
Path A often increases the number and severity of design rules because the process is more sensitive to layout context. Designers may face stricter constraints on minimum spacing, line ends, jogs, vias, and local density patterns. These rules are not aesthetic; they exist to preserve yield. But they can reduce the theoretical density benefit of the node because routing becomes harder and whitespace grows.

Path B can reduce some multipatterning-driven constraints, but it may introduce its own rules tied to the new imaging regime and process integration. Either way, the "node" becomes partly defined by its rulebook, not by its transistor.

Cell libraries and standard-cell architecture
Standard-cell design becomes a negotiation between transistor capability, routing feasibility, and patterning restrictions. If pattern splitting is heavy, cells may need to be more regular and less aggressive, potentially reducing density or performance. If High-NA simplifies some layers, cell architects may regain flexibility—but only if other constraints (contacts, power rails, interconnect) don't dominate.

SRAM scaling
SRAM is often the canary in the scaling coal mine. Logic can sometimes be massaged with design tricks and margining, but SRAM arrays expose systematic patterning and variability

issues. Many advanced nodes are judged by whether SRAM bitcells can shrink while maintaining adequate yield and stability. Path A's overlay and stochastic challenges can make SRAM scaling harder, forcing conservative bitcell designs. Path B may help on some patterning dimensions, but SRAM is also sensitive to variability and contact resistance, which are not solved by lithography alone.

PPA realism

PPA (performance, power, area) gains are commonly advertised, but they can be illusory if design rules and routing penalties erode area gains, or if interconnect RC and contacts limit performance. The lithography path influences how "real" PPA gains are in practice. Path A can deliver device-level gains that are partially cancelled by complexity-driven design penalties. Path B can restore some area and cycle-time benefits but can introduce ramp and availability constraints that delay the moment when those benefits become commercially accessible.

The practical conclusion is that "extend low-NA vs jump to High-NA" is also a choice about design philosophy. Path A tends to force more restrictive, regularized design in exchange for staying on a mature tool platform. Path B tends to promise more patterning headroom in exchange for absorbing platform transition risk. In both cases, the ultimate test is not what is possible in a lab. It is what is deliverable in a design kit that thousands of engineers can use without falling into yield traps.

In the below-2 nm era, scaling strategy is lithography strategy, and lithography strategy becomes design strategy. The foundry's choice determines not only what the factory can print, but what the designer is allowed to imagine.

6. GAA Is Necessary but Not Sufficient: The Full Stack of "Below 2 nm"

Gate-all-around nanosheets are the headline because they are the visible transistor transition. But the sub-2 nm era is defined by a more uncomfortable truth: the transistor is no longer the single dominant limiter. GAA improves electrostatic control and creates new device-sizing knobs, but those gains can be neutralized—or even reversed—by interconnect penalties, power delivery constraints, stochastic defect behavior, and the sheer difficulty of maintaining tight control over a long and fragile manufacturing flow. This is why the correct mental model for "below 2 nm" is a full-stack convergence problem. You can have an excellent device and still ship an inferior product if the wiring, power, variability, and packaging realities do not cooperate.

In practical terms, the "below 2 nm" stack has two dominant themes. First, performance is increasingly wire-limited, not device-limited. Second, economics and reliability are increasingly

governed by control loops: metrology, inspection, feedback, and yield learning. GAA helps, but it does not remove those constraints. It simply enables the next stage of confronting them.

6.1 Interconnect and BEOL limits: resistance, capacitance, and routing congestion

The back-end-of-line (BEOL)—the metal layers that connect transistors—has become one of the least forgiving aspects of scaling. For many workloads and many designs, the performance and power you feel at the chip level is a function of how quickly signals move through wires and how much energy is lost charging and discharging capacitive loads, not how fast an isolated transistor can switch.

Three interconnect pressures intensify below 2 nm-class.

Resistance rises as wires shrink.
As cross-sectional area shrinks, resistance increases. That is simple geometry, but the impact is amplified by materials effects at tiny scales: electron scattering at surfaces and grain boundaries increases effective resistivity. The result is that narrower wires can be disproportionately "worse" than naive scaling predicts. Higher resistance increases delay, increases IR drop, and increases power dissipation in the interconnect itself.

Capacitance does not scale away cleanly.
Even when wire dimensions shrink, spacing may not shrink proportionally because spacing is constrained by patterning and reliability rules. Meanwhile, wires are packed more densely, and coupling between neighboring wires can increase. Capacitance drives dynamic power and can worsen signal integrity. The real danger is that capacitance can remain stubbornly high while resistance rises, producing worse RC delay even as transistors improve.

Routing congestion becomes structural.
As design rules become more restrictive and as standard-cell architectures become more constrained, routing becomes more difficult. This often shows up as congestion hotspots that force detours, additional buffer insertion, and larger cell utilization. The theoretical density advantage of a node can be eroded by routing penalties and whitespace introduced to satisfy design rules. This is where "density" becomes a design-flow reality, not a transistor fact.

A critical consequence follows: below 2 nm, the node's value proposition depends heavily on BEOL strategy, not just on device innovation. If the interconnect stack imposes severe RC penalties, the chip's performance-per-watt gains can underdeliver relative to transistor-level claims. This is one reason advanced nodes increasingly need system-level solutions—like architectural changes and packaging strategies—to keep net gains positive.

6.2 Power delivery transitions: why backside power appears now

Power delivery has moved from being an engineering detail to being a node-defining challenge. At older nodes, designers could often treat power grids as a scaling nuisance. At sub-2 nm, power delivery competes directly with signal routing for scarce metal resources, while current demands and power density rise. This drives the industry toward new power delivery strategies, including backside power delivery.

The basic problem has two coupled parts:

Power rails consume routing resources.
Frontside power delivery uses valuable metal layers that could otherwise route signals. As standard cells get smaller and routing becomes tighter, dedicating significant routing capacity to power becomes increasingly expensive. Signal routing suffers, increasing congestion and delay, and the density gains you paid for can be lost.

IR drop and electromigration margins tighten.
As currents remain high and wires narrow, the voltage drop across the power network becomes more severe. IR drop reduces effective voltage at transistors, hurting performance and increasing the risk of timing failures or unstable operation. Meanwhile, current density increases stress, raising electromigration risk. Reliability becomes more difficult to guarantee, especially for high-performance products.

Backside power delivery appears as a response because it offers a compelling architectural separation: move bulk power delivery away from the frontside signal routing domain. In ideal form, this does two things:

- It frees frontside metal resources for signal routing, reducing congestion and improving routability.
- It shortens and strengthens power delivery paths, potentially reducing IR drop and improving power integrity.

But backside power is not a cosmetic change. It introduces new process complexity, new integration steps, and new failure modes. It demands high precision in wafer thinning, backside via formation, and power distribution integration. The reason it appears now is not fashion; it is because the combined pressure of routing scarcity, power density, and reliability constraints is reaching a point where incremental tuning of frontside power is no longer sufficient for the most aggressive nodes and products.

In the sub-2 nm regime, power delivery transitions should be interpreted as a sign that scaling has become a system architecture problem. The node is no longer just “smaller transistors.” It is a reorganization of where power and signals live.

6.3 Process control: metrology, inspection, and stochastic defect management

If lithography is the governor, process control is the nervous system. Below 2 nm, the tolerance stack is so tight that manufacturing without aggressive measurement and feedback becomes blind flight. The factory must continuously measure, infer, and correct.

Three control domains are especially critical.

Metrology for dimensional control and overlay.
Variations in critical dimensions and layer-to-layer alignment can destroy yield or degrade performance bins. Metrology must not only be accurate; it must be fast and sufficiently sampled to detect drifts before they become catastrophic. Below 2 nm, the factory cannot rely on occasional sampling. It needs a disciplined measurement strategy integrated into production rhythms.

Inspection for defect discovery and suppression.
Inspection finds the defect modes that dominate yield. But inspection itself becomes challenging because defects can be extremely small and because stochastic defects can be rare yet devastating. The goal is not to eliminate all defects—impossible at any practical cost—but to identify the dominant defect sources and reduce them fast enough that yield learning stays ahead of product timelines.

Stochastic defect management as a first-class discipline.
EUV introduces stochastic behavior driven by photon statistics and resist chemistry. At tiny features, rare events become inevitable. Managing this is not just "improve resist." It is a full-stack strategy:

- Adjust dose and imaging conditions to reduce stochastic variability, balanced against throughput and resist blur.
- Use computational correction and pattern restrictions to avoid high-risk geometries.
- Employ inspection and data analytics to detect stochastic-driven failure patterns.
- Establish design rules that trade small density gains for large yield stability gains where needed.

The key insight is that below 2 nm, yield learning is increasingly a data problem. The factory must behave like a closed-loop control system with high-quality sensors and fast corrective action. This also explains why scaling is expensive: you are not only shrinking features; you are building a more precise measurement-and-control machine around the shrink.

6.4 Packaging and system scaling: advanced packaging as the “other half” of node leadership

As interconnect and power constraints intensify and as cost-per-good-die becomes punitive for large monolithic dies, packaging becomes the second half of advanced-node competitiveness. Advanced packaging is not merely an afterthought to protect the die; it is a scaling strategy that can substitute for some forms of lithographic scaling.

The core move is heterogeneous integration: instead of placing all functions on the most advanced node, you partition the system.

- Put the compute logic on an advanced node where performance-per-watt matters most.
- Put large SRAM blocks, I/O, analog, or other area-heavy structures on a more mature node where cost and yield are better.
- Connect the parts with high-bandwidth, low-latency interconnects through advanced packaging.

This approach attacks several constraints at once:

Yield and economics improve via smaller dies.
Smaller dies generally yield better because defect probability scales with area. Partitioning can convert one fragile, expensive die into multiple more yield-friendly components, improving cost-per-good-system.

Scaling becomes system-level.
If interconnect within a die is becoming a bottleneck, moving some connectivity to a package-level interconnect with different physical constraints can help. Packaging can offer new routing resources and new ways to manage power and thermal design.

Time-to-market and risk can improve.
You can iterate one chiplet without respinning the entire system. You can reuse chiplets across product generations. This creates a form of modularity that can be strategically valuable in an era where masks and respins are extremely expensive.

However, advanced packaging introduces its own constraints: thermal density, mechanical stress, yield at the package level, test complexity, and new supply chain dependencies. It does not eliminate complexity; it relocates it. Still, the net effect is often positive because the cost and yield penalties of monolithic scaling at the frontier are so severe.

In this sense, node leadership below 2 nm is increasingly inseparable from packaging leadership. A foundry (or an ecosystem) that can offer a compelling advanced packaging roadmap can capture system-level wins even if its raw transistor scaling is not the most aggressive. Packaging becomes a competitive weapon because it allows customers to capture much of the benefit of

advanced nodes without paying the full risk and cost of putting everything on the newest process.

The synthesis is straightforward: GAA enables the next stage of transistor scaling, but the sub-2 nm era is defined by the full stack. Interconnect and BEOL can erase device gains, power delivery can force architectural transitions, process control determines whether yields become stable, and packaging provides the escape hatch that turns scaling from a die-level problem into a system-level design space. In this regime, the winners will be those who treat "below 2 nm" not as a number but as a coordinated engineering discipline spanning silicon, tools, data, and the package.

7. Competitive Dynamics: Samsung, TSMC, Intel Through the ASML Lens

Seen through the ASML lens, competition at the leading edge is less about who has the smallest nominal node and more about who aligns most effectively with the constraints and opportunities of the lithography platform. EUV—low-NA today, High-NA tomorrow—acts as a common gravitational field. All three major players must orbit it, but they choose different trajectories. Those choices reveal their priorities, risk tolerance, and narratives to customers. The result is not a single "race," but a set of overlapping strategies that trade time, capital, learning risk, and credibility in different ways.

7.1 Differentiation strategies: platform timing, product anchoring, and customer signaling

Each major player differentiates not only through technology, but through *how* and *when* it commits to platform transitions—and how loudly it signals those commitments.

One strategy emphasizes **platform timing**. A company may choose to adopt a new lithography capability early to signal long-term leadership and technological ambition. Early adoption communicates confidence in internal engineering capacity and a willingness to absorb learning costs. It also sends a message to the ecosystem: align with us early if you want a seat at the next platform.

Another strategy emphasizes **product anchoring**. Rather than lead with abstract node claims, a company anchors its node story to specific, recognizable products. This converts uncertainty into narrative clarity: customers and observers can point to a chip and say, "that exists." Product anchoring reassures customers who care more about schedules and deliverables than about theoretical leadership.

A third strategy emphasizes **customer signaling**. This includes selective disclosure of yield progress, careful language around risk production versus volume, and deliberate ambiguity about internal details. Some foundries cultivate an image of conservative credibility: fewer bold

claims, more emphasis on reliability. Others lean into assertive signaling to shape market expectations and investor sentiment.

ASML sits beneath all of these strategies. Platform timing decisions implicitly assume something about ASML's readiness and support. Product anchoring assumes sufficient scanner availability and uptime. Customer signaling often encodes beliefs about how far the lithography platform can be pushed without visible failure. The differentiation is real, but it is bounded by a shared dependency.

7.2 The adoption fork: why one leader may delay High-NA while another pulls it in

The fork between extending low-NA EUV and adopting High-NA EUV early is the clearest place where competitive philosophies diverge.

A leader may **delay High-NA** for several rational reasons. Low-NA EUV may still offer sufficient headroom for the next node if multipatterning and computational correction can deliver acceptable yields and costs. Delaying avoids early platform risk, preserves capital flexibility, and allows the ecosystem (resists, masks, inspection, EDA) to mature further. For a foundry that prioritizes predictable ramps and high-volume reliability, this conservative posture can be strategically coherent—even if it concedes early "headline leadership."

Another leader may **pull High-NA in early** to regain patterning margin and simplify the most challenging layers. This can reduce multipatterning complexity, shorten cycle times for certain layers, and create a narrative of architectural leadership. Early adoption can also be a way to reset the competitive field: if others are optimized for low-NA extension, a platform jump can invalidate some of that optimization and create differentiation.

The choice is also shaped by **internal structure**. An integrated device manufacturer with tight coupling between design and manufacturing may tolerate higher early risk because it can co-optimize aggressively and accept internal disruption. A pure-play foundry serving many external customers may place a higher premium on stability and broad design enablement.

From the ASML perspective, both choices reinforce dependence. Delayers drive higher lithography intensity on low-NA tools; early adopters validate the High-NA roadmap and pull the ecosystem forward. The divergence affects *when* ASML's next platform monetizes, not *whether* it matters.

7.3 Geopolitics and supply chain realities: export controls, capacity siting, and resilience

Below-2 nm competition unfolds under geopolitical constraints that directly intersect with lithography. Export controls shape who can receive the most advanced tools, when they can receive them, and under what conditions. These controls do not merely affect national champions; they influence global capacity planning and risk management.

Capacity siting has become strategic rather than purely economic. Advanced fabs are now distributed across regions to satisfy political, security, and resilience goals. This dispersion complicates ramp logistics: duplicating EUV capacity across sites requires more tools, more trained personnel, and more service coverage. ASML's role expands here, because service quality, spare availability, and field engineering depth become as important as tool capability.

Resilience has also become a competitive dimension. Customers increasingly care about supply continuity under stress scenarios—geopolitical shocks, trade restrictions, or natural disruptions. A foundry's ability to assure continuity is partly a function of its lithography footprint and its relationship with the sole EUV supplier. Redundancy is expensive when the bottleneck tool is unique, but the lack of redundancy is a business risk.

These realities mean that "leadership" cannot be assessed in isolation from political context. A node that is technically superior but geopolitically constrained may be less attractive than a slightly less aggressive node with assured access and stable delivery. Lithography is where these constraints crystallize, because it is both technologically indispensable and politically sensitive.

7.4 What "leadership" becomes: reliability, capacity, and ecosystem—beyond first claims

In the early days of a new node, leadership is often narrated as "first to announce" or "first to demonstrate silicon." In the below-2 nm regime, that definition becomes dangerously incomplete. Leadership evolves into a composite of attributes that are harder to advertise but easier to feel.

Reliability becomes paramount. Customers value a node that behaves the same way across wafers, lots, and quarters. Reliability includes yield stability, parametric consistency, and predictable performance bins. It also includes the absence of late-stage surprises that force redesigns or respins.

Capacity becomes strategic. It is not enough to have a working process; you must be able to scale it. Capacity depends on tool availability, uptime, staffing, and the ability to replicate success across fabs. Because lithography is the gating step, capacity leadership is inseparable from access to, and mastery of, ASML's tools.

Ecosystem strength becomes decisive. A leading node supports a broad range of products with mature IP, robust EDA flows, packaging options, and test solutions. It allows customers to innovate without constantly fighting the process. Ecosystem maturity often lags transistor invention, but it ultimately determines commercial impact.

Through the ASML lens, leadership is therefore not a moment; it is a sustained condition. It is the ability to translate lithography capability into manufacturable reality at scale, under economic and geopolitical constraints, while enabling customers to succeed repeatedly.

Samsung, TSMC, and Intel pursue this condition along different paths, but all are bounded by the same lithographic physics and the same platform supplier.

The deeper lesson is that at the frontier, competition is less about outpacing rivals by a single metric and more about *outlasting complexity*. In that contest, lithography is not merely a tool—it is the tempo-setter that determines who can keep moving when the margins disappear.

8. Implications for Customers: What to Ask Before Betting a Product on a Node

For a customer, choosing a leading-edge node is not a referendum on technological purity. It is a commitment to a schedule, a cost structure, and a risk profile. The below-2 nm era intensifies this because the gap between "silicon exists" and "profitable, repeatable volume" can be large, and because the penalty for surprises is harsher: mask costs are higher, design rules are more constraining, and cycle time is longer. A customer that treats node selection as a branding decision can lose a year. A customer that treats node selection as an evidence-based risk decision can buy optionality, protect the schedule, and still capture most of the performance-per-watt upside.

This section frames the node decision as due diligence. The key is to ask questions that convert marketing into measurable commitments. The customer cannot demand to see every proprietary chart, but it can demand enough structure to judge credibility: stability of the process window, maturity of the design kit, cycle-time realism, defectivity trends, and whether the foundry's roadmap is internally coherent rather than aspirational.

8.1 The customer due diligence checklist: yield, cycle time, defectivity, and roadmap credibility

A disciplined checklist is the customer's main defense against being seduced by node labels. The checklist should be organized around the variables that actually determine whether the product launches on time and hits margin targets.

Yield maturity and yield predictability
Do not ask only "what is your yield?" Ask questions that expose whether yield is stable and generalizable.

- What is the yield distribution over time (trend and variance), not just a single point?
- How does yield scale with die area? What is the practical die-size ceiling for acceptable yield today?
- How sensitive is yield to layout context? Are there known "yield traps" or forbidden structures?

- What is the split between catastrophic yield loss (dead die) and parametric yield loss (dies that work but bin low)?
- How many tapeouts have successfully gone through this node, and how diverse are they?

Cycle time and schedule realism
Cycle time is often the silent killer. In advanced nodes, long cycle time slows learning loops and extends debug timelines.

- What is the current wafer cycle time for the critical layers and for the full flow?
- How variable is cycle time under load? What happens when the fab is at high utilization?
- What is the rework rate for key layers, and how does rework affect cycle time predictability?
- What are the typical turnaround times for engineering lots versus production lots?

Defectivity and defect learning
Customers may not get full Pareto charts, but they can demand enough to assess whether defect learning is mature.

- What are the dominant defect modes today, and how have they changed over the last quarters?
- What is the foundry's inspection and metrology coverage strategy?
- How quickly can the fab identify and suppress a new defect mode? What is the typical "time to containment"?
- Are there known stochastic defect issues on specific layers, and how are they managed?

Roadmap credibility and internal coherence
Foundry roadmaps can be ambitious; the customer's job is to evaluate coherence and deliverability.

- What is the timeline for risk production versus volume, and what are the gating criteria?
- How does the foundry define "ready" for external customers?
- Are the announced enhancements (yield, performance, density) tied to specific process changes with plausible timelines?
- Does the foundry have a credible capacity plan aligned with tool availability?

A good litmus test: a credible roadmap describes constraints and tradeoffs, not just advantages. A roadmap that sounds too smooth is often masking schedule risk.

8.2 Cost realism: masks, tool time, and the hidden tax of complexity

Below-2 nm economics punish naïve budgeting. Customers often underestimate cost because they focus on wafer price and forget the multiplicative costs: masks, EDA effort, design closure, validation, respins, and the opportunity cost of delay. Cost realism requires treating the node as a total program cost.

Masks as a direct cost and a schedule constraint
Advanced nodes require large mask sets, and the cost of a single mask set can be extraordinary. But the deeper issue is not just cost; it is time. Mask turnaround can become a schedule gate, and additional masks introduced by multipatterning can amplify the problem.

Customers should ask:

- How many masks does the full flow require, and how many are on the critical path for early silicon?
- How does the foundry support mask schedule acceleration?
- What is the expected mask respin rate for first-time customers on this node?

Tool time as "invisible pricing"
Even if wafer pricing is negotiated, the real scarcity is often tool time on critical lithography layers. If the process requires more exposures (pattern splitting), the same wafer output consumes more scanner time. That scarcity can show up as queueing, longer cycle times, and priority allocation.

Customers should ask:

- Is there guaranteed access to key layers during ramp?
- What are the lot priority policies, and do anchor customers receive preferential treatment?
- How does the foundry manage tool downtime without disrupting customer schedules?

The hidden tax of complexity
Complexity is not merely "more steps." It increases the probability of late surprises: a marginal patterning issue that becomes catastrophic at scale, a new yield trap that appears when routing density changes, or a parametric distribution that drifts under high-volume conditions.

Customers should budget explicitly for:

- Additional design iterations (layout rework to satisfy evolving rules)
- Additional verification effort (corner analysis, EM/IR, thermal, reliability)
- Potential extra silicon spins (and the schedule margin to survive them)
- Extra validation time due to binning variability and power/performance tuning

A useful rule: when process complexity rises, cost uncertainty rises. If a foundry is "stretching" an existing platform, the customer should treat complexity tax as a first-class line item, not a rounding error.

8.3 Risk posture: dual sourcing, design portability, and contingency planning

In the below-2 nm era, the biggest program failure mode is not that the chip fails electrically. It is that the program fails to launch on time at an acceptable cost. Risk posture is how customers protect themselves against schedule volatility and yield uncertainty.

Dual sourcing
True dual sourcing at the leading edge is hard, but partial forms are possible.

- Dual-source at the product portfolio level: one flagship product on the frontier node, another product on a mature node to stabilize revenue.
- Dual-source at the foundry level where feasible: keep a fallback plan on a more mature node at a second supplier, even if performance differs.
- Dual-source through packaging strategy: design modular chiplets so that not all functionality depends on the most advanced node.

The key is to avoid a single point of failure: one node, one foundry, one schedule.

Design portability and "escape velocity"
Portability is not "compile once, run anywhere." Each node's design rules and libraries can be deeply specific. But customers can still design for escape.

- Use modular architectures that can tolerate changes in frequency or power targets.
- Avoid overly node-specific micro-optimizations until the process is mature.
- Maintain a design methodology that can be retargeted with reasonable effort.
- Preserve optionality in IP choices: do not lock into a fragile IP stack that depends on one foundry's unique assumptions.

Contingency planning

A credible plan assumes things will go wrong. The question is whether the program has pre-decided what it will do.

- If yield is below target at date X, do you reduce die size, adjust SKU strategy, or switch packaging?
- If frequency bins are lower than expected, do you adjust marketing claims and product segmentation?
- If schedule slips, what revenue bridge exists?

Customers should treat these not as pessimism but as engineering maturity: you are not predicting failure; you are preventing failure from becoming catastrophic.

8.4 The "good enough" node: when a mature node can beat a bleeding-edge node

There is a recurring truth in semiconductors: the best node is not always the smallest. A mature node can win when it offers better cost-per-good-die, better predictability, and faster time-to-market. The below-2 nm era makes this even more common because advanced nodes can be economically fragile early on.

A mature node can beat a bleeding-edge node when:

- **Time-to-market dominates.** If shipping earlier wins market share, a stable node can outperform a late, theoretically superior node.
- **Yield and binning matter more than peak performance.** For many products, predictable bins and supply are more valuable than absolute maximum frequency.
- **The design is not scaling-limited.** Some workloads are memory-limited, I/O-limited, or power-limited in ways that do not benefit proportionally from the newest node.
- **Packaging and architecture can substitute for scaling.** Chiplets, advanced packaging, and heterogeneous integration can capture much of the system benefit without placing the entire design on the newest node.
- **Total program economics rule.** If mask costs, iteration cycles, and risk margins overwhelm the performance benefit, the "best" node is the one that ships profitably.

The customer's disciplined posture is therefore: do not fetishize the frontier. Use it when it is necessary for a clear product advantage and when the schedule and yield risks are priced and managed. Otherwise, treat the frontier as an option you exercise only when it is truly worth the premium.

In the below-2 nm regime, the most sophisticated customers will not be those who always chase the newest node. They will be those who can quantify when the newest node is justified, when it is premature, and how to preserve optionality so that no single node decision can sink the product.

9. A Disciplined Way to Talk About "Below 2 nm" Without Hype

"Below 2 nm" has become a rhetorical accelerant. It compresses a complicated technical landscape into a single number that feels like a finish line, even though the real work is a multi-year convergence of patterning, yield learning, design enablement, and economics. In this environment, hype is not just annoying; it is operationally harmful. It can cause customers to bet products on immature platforms, cause investors to misread readiness, and cause public discourse to confuse branding with capability.

The remedy is not cynicism. The remedy is disciplined language: vocabulary that separates types of claims, an evidence hierarchy that distinguishes signal from narrative, and a shared norm that precision is not pedantry but an ethical requirement. At the frontier, vague language is not "harmless simplification." It is often an implicit lie, even when nobody intends to lie, because it invites the listener to infer readiness that has not been demonstrated.

9.1 A vocabulary of claims: "branded," "qualified," "risk," "HVM," "profitable HVM"

A useful vocabulary does one thing: it forces the speaker to say what kind of claim they are making. The following terms form a minimal, practical ladder. The point is not to impose a universal standard. The point is to stop mixing categories.

Branded
A node is "branded" when it has a name, a public identity, and a place on a roadmap. "Branded" is not meaningless. Branding coordinates the ecosystem: it tells EDA teams, IP vendors, and customers what to prepare for. But branding is not proof of manufacturability. It is a declaration of intent.

Qualified
A node is "qualified" when the foundry has internally defined criteria it believes are sufficient for a certain class of production use. Qualification can refer to basic process stability, reliability targets, or a subset of layers being mature. Qualification can be real and still be narrow. The key question is always: qualified for what product types and what die sizes?

Risk
A node is in "risk" when real customers can tape out meaningful designs and expect silicon that is not merely a science experiment. Risk implies commitment, but it also implies recognized

uncertainty: evolving design rules, shifting process windows, yield volatility, and the possibility that early customers will fund the learning curve through schedule pain.

HVM (High-Volume Manufacturing)
HVM is the claim that the node is producing at scale with repeatability. In practice, HVM implies stable yields, predictable cycle time, and the ability to meet sustained customer demand. HVM is not "we can run wafers." It is "we can run wafers predictably at commercially meaningful volumes."

Profitable HVM
This is the term that most public narratives avoid, because it forces economic clarity. Profitable HVM means the cost-per-good-die is competitive enough that the foundry and the customer can make money without heroic exceptions. It implies that yields are not only high but stable, that bins are predictable, and that the node's complexity does not impose hidden costs that erase the performance-per-watt advantage.

This ladder solves a common confusion: people treat a "branded" node as if it were HVM, or treat "risk" as if it were profitable HVM. The disciplined speaker always states where on the ladder they are talking.

A simple norm follows: if you do not know, say "branded" or "planned," not "shipping."

9.2 Evidence hierarchy: what counts as signal versus PR

In the below-2 nm era, the information environment is adversarial—not because people are malicious, but because incentives reward optimistic narratives. The only defense is an evidence hierarchy that prioritizes what is hard to fake.

A practical hierarchy looks like this, from strongest to weakest:

Tier 1: Market behavior under constraint
What chips are actually shipping? At what volumes? With what customer commitments? Behavior under constraint is the hardest thing to fake because it involves supply chains, contracts, and real money.

Tier 2: Repeatable manufacturing indicators
These are signals that imply stable process operation: multiple products on the node, diverse design styles, predictable launch schedules, and credible statements about capacity planning. You may not see the raw yield number, but you can often see the pattern of operational maturity.

Tier 3: Ecosystem maturity
Availability and stability of PDKs, standard cell libraries, SRAM compilers, analog IP, and EDA sign-off flows. Ecosystem maturity often lags technical demonstration, but when it is present, it is a strong signal that a node is becoming real for external customers.

Tier 4: Demonstrated silicon
"First silicon" is a necessary milestone, but it is not sufficient. A single demo chip can hide enormous amounts of engineering heroism, conservative design, or special treatment in the fab. Demonstrated silicon is signal, but modest signal.

Tier 5: Announcements and roadmap slides
These are the easiest to produce and the easiest to misinterpret. They are still useful for understanding intent and strategic direction, but they should never be treated as evidence of manufacturability.

Under this hierarchy, PR is not "false." It is simply low-grade evidence. It tells you what a company wants you to believe, not what has been forced into truth by the factory.

The disciplined analyst uses the hierarchy to ask: what kind of evidence am I looking at, and what conclusions does that evidence actually justify?

9.3 The public's trap: why precision language is now a form of technical ethics

The public trap is not that people are ignorant. The trap is that ordinary language invites ordinary inferences. When a headline says "1.4 nm," a reasonable person hears "smaller than 2 nm," and then hears "more advanced," and then hears "ready." This cascade is almost automatic. The problem is that each step is less justified than it feels.

Precision language is therefore a form of technical ethics because it protects others from predictable misunderstanding. In this era, ambiguity does not remain neutral; it is almost always interpreted in the most optimistic direction. That means a speaker who uses vague language in public is not merely simplifying; they are shaping decisions.

This matters in at least three stakeholder domains:

Customers can be induced to bet products on immature nodes.
A customer that mistakes "risk" for "HVM" can discover too late that the schedule is not recoverable without respins or product segmentation.

Investors and policymakers can misallocate capital and incentives.
If "node leadership" is treated as a binary, it can drive funding toward visible announcements rather than toward the less glamorous work of yield learning, workforce development, and resilient supply chains.

The technical community can lose trust.
When the public repeatedly experiences that “breakthrough” claims are followed by delays and caveats, credibility erodes—even for genuinely impressive progress.

So precision language is not pedantry. It is stewardship. It treats the listener as someone who deserves to know what is real, what is planned, and what is uncertain.

A simple ethical rule fits the below-2 nm era:

- Do not let a node label do work that only yield and volume can do.
- Do not let a demo do work that only repeatability can do.
- Do not let an announcement do work that only shipping can do.

In this disciplined frame, “below 2 nm” becomes a meaningful phrase again. It stops being a magic number and becomes what it actually is: a name for a regime where the truth is manufactured through ruthless control of variation, defects, and economics—and where careful language is part of that manufacturing process because it prevents narrative from outrunning reality.

10. Conclusion: The Sub-2 nm Era as Lithography-Governed Systems Engineering

The sub-2 nm era is best understood as a change in what it means to make progress. The industry is not simply shrinking a feature and collecting an automatic reward. It is negotiating a tightening tolerance stack in which pattern fidelity, overlay, stochastic defect rates, variability, and economic viability are coupled so tightly that no single subsystem can “win” alone. In this regime, the most important innovations are often not the most visible ones. They are the innovations that make the factory behave like a stable instrument: repeatable, measurable, correctable, and scalable.

If the previous era celebrated geometric shrink as a headline, this era celebrates *convergence*: getting transistor architecture, lithography capability, process control, interconnect, power delivery, and design enablement to land on the same solution rather than on competing local optima. The result is that “leadership” becomes less about being first to name a node and more about being first to make the node boring—boring in the sense that customers can bet products on it without gambling the company.

10.1 Restate the thesis: below 2 nm is a regime, not a number

The thesis of this paper is that “below 2 nm” is not a literal measurement and not a single finish line. It is a regime—a new operating condition for the entire semiconductor stack. In this regime:

- Node labels function primarily as generational identifiers, not as precise geometry.
- Yield learning and cost-per-good-die become the defining scoreboard.
- Lithography capacity and capability shape what is economically printable, and therefore what is designable.
- The bottlenecks move: if you relieve one constraint (transistor control), another (interconnect, power delivery, stochastic defects, overlay) rises in prominence.
- Scaling becomes a coupled systems problem, where the “best” solution is often the one that balances constraints rather than maximizes any single metric.

The practical implication is that conversation must shift. Instead of asking “who is below 2 nm?” we should ask “who can sustain a profitable, repeatable manufacturing platform in the sub-2 nm regime, at the volumes and yields that matter?”

10.2 Where Samsung fits: GAA ambition plus the real test of yield credibility

Samsung’s role in this regime is defined by an ambitious bet: make GAA nanosheets a stable platform early enough to convert learning into a durable advantage. The strategic appeal is clear. GAA offers stronger electrostatic control and new degrees of freedom in device sizing, which are necessary ingredients for continuing scaling. But the decisive question is not whether Samsung can build the transistor. The decisive question is whether Samsung can generalize the process across diverse customer designs and die sizes with a credible ramp.

In other words, Samsung’s success will be measured in credibility, not in novelty.

Credibility has specific content in the sub-2 nm regime:

- Yields that improve and stabilize over time, rather than oscillate by design style.
- Design rules that are restrictive for good reasons but not so restrictive that they erase density gains.
- Cycle time and throughput that support real product schedules.
- Evidence of ecosystem maturity: PDK stability, IP readiness, SRAM viability, and predictable sign-off.

If Samsung can deliver that package, its early GAA bet becomes more than a technology story. It becomes a platform story: a claim that Samsung can be a stable destination for advanced-node product roadmaps. If it cannot, the same early bet becomes a cautionary tale about how sub-2 nm punishes premature platform transitions.

The point is not to predict the outcome with certainty. The point is to clarify what the outcome *means*. In this regime, Samsung's ambition is admirable, but the real test is whether the node becomes boring enough that external customers treat it as schedule-safe.

10.3 Where ASML sits: platform control, adoption timing, and compounding leverage

ASML's position in the sub-2 nm era is structurally central because lithography is the hidden governor: it determines what patterns are feasible, how many steps are required, and how fast the factory can learn. EUV is not simply a tool category; it is the platform that shapes the entire economic and technical envelope of the leading edge.

ASML's leverage manifests in three ways.

First is **platform control**. As the sole supplier of EUV scanners, ASML anchors the advanced-node ecosystem. This does not mean ASML "controls" outcomes in a simplistic way, but it does mean that the timing and capability of its platform heavily influence the feasible choices of every leading-edge manufacturer.

Second is **adoption timing**. The fork between extending low-NA EUV and adopting High-NA EUV early is one of the most consequential strategic decisions in the sub-2 nm era. But regardless of which path a foundry chooses, ASML remains central: extending low-NA often increases lithography intensity (more exposures, more tool time), while adopting High-NA pulls the ecosystem into a new platform with new dependencies. Either way, the industry's progress remains bounded by ASML's tool roadmap, tool availability, and operational maturity.

Third is **compounding leverage through the installed base**. At the leading edge, uptime and service are not operational details; they are competitive variables. A large installed base, extensive field data, and the ability to upgrade productivity and reliability create compounding advantage. In a high-intensity lithography regime, marginal gains in throughput or uptime can translate into significant effective capacity, faster learning loops, and earlier profitability for customers. That makes the supplier's operational excellence as important as its headline capability.

So "where this leaves ASML" is not a question of relevance. It is a question of how the industry negotiates the platform transition curve: whether High-NA adoption accelerates quickly or whether low-NA is stretched further. In both cases, ASML becomes more central because the sub-2 nm regime is defined by lithography-constrained systems engineering.

10.4 Forward work: case studies, comparative timelines, and a disciplined Reference section

The forward work for this paper is to turn the conceptual map into concrete, testable analysis. Three extensions are especially valuable.

Case studies
Select a small number of representative products and trace what the node choice meant in practice: schedule, yield evolution, performance bins, and cost outcomes. A case study should not merely list claims; it should identify the constraint that dominated (overlay, stochastic defects, contacts, interconnect, cycle time) and show how the program responded.

Comparative timelines
Build a timeline that separates branding, risk production, HVM, and profitable HVM for each major player. The goal is not to "rank" companies but to clarify how long the gap is between categories and how that gap changes over time. The timeline should explicitly incorporate lithography platform milestones, because those milestones often explain why a ramp accelerates or stalls.

A disciplined Reference section
Because this paper avoids in-line references by design, the Reference section should be structured to preserve credibility: primary sources first (company technology symposium materials, official roadmap statements, credible industry reporting), then secondary commentary. The references should also be organized around the paper's central claims: (a) node naming and the marketing/measurement disconnect, (b) GAA transition details, (c) low-NA vs High-NA EUV platform realities, and (d) customer-facing manufacturability indicators.

A good final step is to formalize the "disciplined vocabulary" introduced in Section 9 into a one-page appendix: a glossary and an evidence hierarchy checklist. That would allow readers to apply the framework beyond this specific moment in the roadmap cycle.

The sub-2 nm era will produce many headlines. This paper's goal has been to supply something more durable: a way to interpret those headlines without being fooled, and a way to see the real story beneath the node names. The real story is that progress now requires system-level convergence under lithography-governed constraints. The winners will not be those who announce first. They will be those who manufacture truth—repeatably, profitably, and at scale.

References

1. Samsung Electronics (Samsung Foundry). “Samsung Foundry Showcases Advanced Technology and Solutions.” Tech blog, Feb 19, 2024. Samsung Semiconductor Global
2. Samsung Electronics. “Samsung Showcases AI-Era Vision and Latest Foundry Technologies at SFF 2024.” News release, Jun 12, 2024. Samsung Semiconductor Global
3. ASML. “5 things you should know about High NA EUV lithography.” ASML Stories, Jan 25, 2024. ASML
4. ASML. “Intel and ASML strengthen their collaboration to drive High-NA into manufacturing in 2025.” Press release, Jan 19, 2022. ASML
5. ASML. “EUV lithography systems – Products.” Product overview page (includes High-NA platform notes and timeline). ASML
6. TSMC. “A16 Technology.” TSMC logic technology page. TSMC
7. TSMC. “TSMC Celebrates 30th North America Technology Symposium.” Press release, Apr 24, 2024 (A16 and N2P PPA statements). TSMC
8. Intel. “Semiconductor Manufacturing Process | Intel 18A, 3, and 16.” Intel Foundry process overview page. Intel
9. Tom’s Hardware. “Samsung’s new roadmap unveils its 2nm process nodes and outlines backside power delivery plans.” Jun 13, 2024. Tom's Hardware
10. Tom’s Hardware. “Intel installs industry’s first commercial High-NA EUV lithography tool — ASML Twinscan EXE:5200B sets the stage for 14A.” (Published Dec 2025 per article metadata). Tom's Hardware
11. Tom’s Hardware. “TSMC begins quietly volume production of 2nm-class chips.” (Published late Dec 2025 per article metadata). Tom's Hardware
12. Mukesh, S., et al. “A Review of the Gate-All-Around Nanosheet FET Process and Scaling Challenges.” *Electronics* 11(21):3589, 2022. MDPI
13. Wang, M., et al. “A Review of Reliability in Gate-All-Around Nanosheet Transistors.” (Open-access review, 2024). PMC
14. Maas, R., et al. “Stochastics in extreme ultraviolet lithography: investigating the role of shot noise and chemical stochastic effects.” *Journal of Micro/Nanolithography, MEMS, and MOEMS* 17(4), 2018. SPIE Digital Library

15. Mack, C. A. “Shot noise: a 100-year history, with applications to lithography.” (Whitepaper/lecture note, 2018). Lithoguru

16. International Workshop on EUV Lithography. “2017 International Workshop on EUV Lithography — Abstracts (stochastic limits discussion).” Proceedings PDF, 2017. EUV Litho

17. imec. “Introducing 2D-material based devices in the logic scaling roadmap.” Article, Jan 16, 2025. imec

18. Yoon, H. H., et al. “Enabling the Angstrom Era: 2D material-based multi-bridge channel CFET roadmap.” *npj 2D Materials and Applications*, 2025. Nature

19. IRDS (International Roadmap for Devices and Systems). “IRDS—International Roadmap for Devices and Systems: Rebooting Computing (S3S-related publication).” (Roadmap perspective on post-nanosheet / CFET era). ResearchGate

20. BoWonder (IEDM Tutorial material). “Advances in CMOS Technologies and Future Directions…” IEDM 2024 tutorial handout (FinFET, GAA, backside power, CFET topics). iedm24.mapyourshow.com

The author has published over 4,700 AI-assisted papers at: https://www.researchgate.net/profile/Douglas-Youvan/research . Google Scholar has indexed these preprints, and if you want a particular reference, the AI mode of a Google search (Gemini) has efficiently catalogued these preprints.

Below 2 nm
Samsung 2nm GAA, "Angstrom-Class" Branding, and
Where ASML Sits in the New Lithography Regime.
2nm GAA Technology
EUV
ASML
EUV Lithography
1.4nm
A16
18A
ASML